Feasibility of Using a Large Clinical Data Warehouse to Automate the Selection of Diagnostic Cohorts

Reejis Stephen, MBBS[1] Aziz Boxwala, MBBS, PhD[1] Paul Gertman MD[2]
[1]Decision SystemsGroup, Brigham & Women's Hospital, Harvard Medical School, Boston, MA
[2]Eclipsys Corporation, Boston, MA

Data from Clinical Data Warehouses (CDWs) can be used for retrospective studies and for benchmarking. However, automated identification of cases from large datasets containing data items in free text fields is challenging. We developed an algorithm for categorizing pediatric patients presenting with respiratory distress into Bronchiolitis, Bacterial pneumonia and Asthma using clinical variables from a CDW. A feasibility study of this approach indicates that case selection may be automated.

BACKGROUND

The use of CDWs to identify case-cohorts is interesting for comparative benchmarking and retrospective studies. However, automated identification of cases is challenging because some of the data required to evaluate the diagnostic criteria are recorded in free-text form in discharge summaries and other clinical notes. The extraction of data from free text would require the use of sophisticated natural language parsing tools. Prior to expending significant effort in implementing NLP tools, we conducted a feasibility test to assess whether case-cohorts of diagnostic groupings could be determined from data in a CDW.

METHODOLOGY

An algorithm was developed to classify cases of respiratory distress into 5 diagnostic categories: Bronchiolitis, Bacterial Pneumonia, Bronchial Asthma, Bronchiolitis with Asthma and Bacterial Pneumonia with Asthma. The diagnostic criteria in the algorithm were based on literature reviews and our experience (1,2). The algorithm utilizes ten variables, five of which are recorded in free text notes. The discrete, coded data items used were: Band-form count, absolute neutrophil count, white blood cell count, eosinophil count, immunological testing for RSV antigen. The data items that need to be extracted from the text are the presence of respiratory distress, findings suggestive of consolidation or atelectasis in radiographs, family history of asthma, or recurrent asthma attacks and fever. The data items were abstracted to binary variables using the diagnostic criteria (e.g., high circulating band form count = band form count > 5%). The categorization was performed using logical combinations of these binary variables.

The identification the relevant documents were performed by SQL queries contained string-matching patterns. For example, the presence of respiratory distress was determined from the history and physical examination by the phrase "respiratory distress", and equivalent synonymous terms, or of a high respiratory rate for that age. These documents were manually inspected and appropriate findings and values were extracted.

The algorithm was manually tested on a sample population from an anonymized CDW that contains data from two community hospitals. We determined case selection criteria based on the requirements of the algorithm. The algorithm was applied to the selected cases and output was compared with the discharge diagnosis for each visit obtained from the database. Even though the discharge diagnosis, which, was coded in ICD-9, is considered unreliable, this was the most suitable "gold standard" available for use in this feasibility study. We expect to use clinical experts' interpretations for future studies.

RESULTS

From an initial population of 24,213 pediatric patients, a sample population of 61 was obtained after applying the inclusion and exclusion criteria. The discharge diagnoses correlated well with the algorithm's diagnoses, which identify the individual diagnostic category (as defined above) in 87 percent of the cases.

DISCUSSION

The preliminary tests indicate this approach can be used to automate the selection of cohorts efficiently and reliably as compared to manual chart review. We are testing the approach on a larger dataset for more conclusive results. We are developing NLP tools for this application and will assess differences in performance.

Acknowledgments

Reejis Stephen is supported by a contract from the Eclipsys Corporation.

References

1. Kumar N, Singh N, Locham KK, Garg R, Sarwal D: Clinical Evaluation of Acute Respiratory Distress and Chest Wheezing in Infants. Indian Pediatrics 2002; 39:478-483.
2. Behrman RE, Kliegman RM, Jenson HB, et al, eds. Behrman: Nelson Textbook of Pediatrics, 16th ed. W.B.Saunders Company, 2000.